Interim Report for ONR Grant: "New Neural Algorithms for Self-Organized Learning"

John E. Moody, PI
Yale Department of Computer Science
New Haven, CT 06520-2158

1 Introduction

This interim report describes work completed from December 1988 to November 1991.

The original purpose of the research program funded by this grant was to study self-organized systems which adapt and learn. The originally proposed research fell into two main categories:

- Biological Models of Self-Organization.
- New Self-Organized Learning Systems.

During the period for which this grant was funded, significant progress was made in both areas.

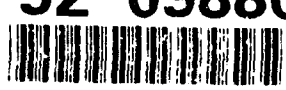
In addition, some new areas related to neural network learning are being explored as an outgrowth of the original proposal. These include:

- Model Selection Techniques.
- Estimating Generalization Performance.

I will divide the discussion of research results into three sections covering work done prior to December 1988, work originally proposed for the period December 1988 to November 1991, and new topics pursued during December 1988 to November 1991. In the discussion below, I will refer to the sections of the original application in which the research was proposed. (The table of contents of the original proposal is included at the end of this document as an appendix.)

Three papers are enclosed which cover work on new learning systems based on localized basis functions as described in the original proposal.

Seven papers (enclosed) were completed during the period 12/88-11/91 under the auspices of this grant. Two papers address the spontaneous development of modular structures (eg. cortical columns) in simple cortical models. One paper addresses the dynamics of lateral interaction networks. Two propose learning rate schedules for fast adaptive k-means clustering and fast stochastic optimization. One described principled architecture selection methods for backprop nets. The last introduced the effective number of parameters $p_{eff}(\lambda)$ and the generalized prediction error (GPE) estimate of generalization performance.



2 Earlier Work: Pre December 1988

In sections 4.1 and 4.2 of the original proposal, we described work which we had begun on several new learning learning systems. This work has since been published in three papers which are included here for completeness: "Fast Learning in Multiresolution Hierarchies", "Fast Learning in Networks of Locally Tuned Processing Units", and "Learning with Localized Receptive Fields".

3 Proposed Work Completed During 12/88 - 11/91

3.1 Spontaneous Development of Modularity

In the original proposal, I described two classes of biological models of self-organization:

- Topographic Models: Dynamics of Self-Organizing Cortical Maps. (See section 3.1 of proposal.)
- Combinatorial Models: Self-organizing Associative Memories. (See section 3.2 of proposal.)

During 1989 and 1990, I worked with a student (Alex Chernjavsky) to develop a detailed network model for the spontaneous development of cortical modules. Our development of this topographic model originally began as an attempt to understand a model of neuronal group formation due to Pearson, Finkel, and Edelman. (See section 3.3 of proposal.)

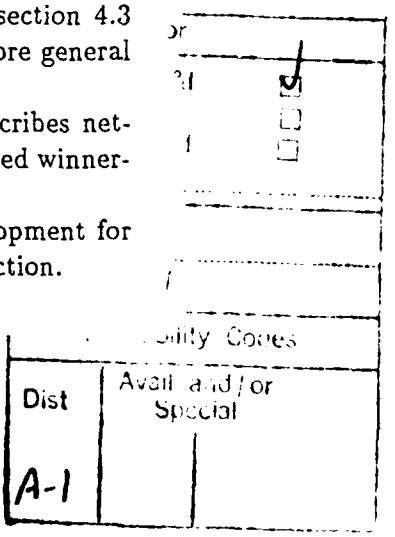
As our work progressed, I realized that we had discovered a generic mechanism for the formation of a wide variety of modular structures in the nervous system, including cortical columns. The results of this work are described at length in Chernjavsky and Moody, "Sponateous Development of Modularity in Simple Cortical Models", Neural Computation, 1990. A less complete description entitled "Note on Development of Modularity in Simple Cortical Models" appeared in **Advances in Neural Information Processing Systems II**, David Touretzky, ed. Morgan Kaufmann, 1990.

3.2 Lateral Interaction Dynamics

Understanding the dynamic behavior of networks is very important for both biological and artificial models. In the original proposal, I described two dynamical network models which compute the 'winner-take-all" and "exp-norm" or "soft-max" functions. (See section 4.3 of proposal.) Since then, I have extended the study of network dynamics to more general network models.

My short conference paper "Dynamics of Lateral Interaction Networks" describes networks which either exhibit "collective excitations" or compute a related "distributed winner-take-all" function.

My discovery of the collective excitation phenomenon was an essential development for understanding the formation of modular structures described in the previous section.



3.3 Fast, Adaptive K-means Clustering

In the original application, we proposed to compare the performance of various numerical techniques for optimizing learning systems. (See section 4.5.) Since the time that was written, we focussed our work on real time learning systems and found that stochastic optimization techniques are required. We also realized that there are not one, but two essential efficiency requirements for practical learning systems: computational efficiency and statistical efficiency.

In a real time context where new data are acquired continuously, statistical efficiency refers to the amount of information which the learning system is able to extract from each new exemplar. The computational cost per exemplar is the number of operations needed to extract the information from each new exemplar. There may often be trade offs between statistical efficiency and computational cost.

A statistically efficient algorithm will converge to a learned solution of a problem after seeing fewer exemplars than will a statistically inefficient algorithm. A computationally efficient algorithm requires relatively few operations per exemplar.

In real time learning systems, the choice of learning rate is a critical factor in determining both statistical and computational efficiency. Conventional neural network learning systems use constant learning rates which are typically tuned by hand. This hand-tuning approach is unsatisfactory and does not generally lead to rapid convergence of on typical learning problems.

We used the classical theory of stochastic approximation to guide us in developing decreasing learning rate schedules which offer much faster convergence in real time contexts. We chose the problem of adaptive k-means clustering or adaptive vector quantization as a test for out methods, because it is one of the most important unsupervised or self-organizing learning algorithms.

In the paper "Fast, Adaptive K-means Clustering" (Chris Darken and John Moody) IJCNN '90 Conference Proceedings, we proposed the new inverse square root learning rate schedule and show that it significantly out-performs all classical or neural net methods which are commonly used.

In the paper "Note on Learning Rate Schedules for Fast Stochastic Optimization" (Darken and Moody, NIPS 90 Proceedings), we proposed and even more powerful class of learning rate schedules called "search-then-converge" schedules which achieve the theoretically optimal asymptotic convergence rates for on-line learning. This work is very general and applies to not only unsupervised learning such as k-means clustering, but also to supervised learning systems, like LMS filters and back-prop networks.

4 Additional Work Completed During 12/88 - 11/91

4.1 Learning Theory: Generalization and Regularization

My recent theoretical work is on generalization and regularization in nonlinear learning systems (preliminary results described in Moody 1991, proceedings of the IEEE neural networks and signal processing conference). The key results include: 1) an expression which relates the expected test and training set errors for nonlinear learning systems which may include regularizers, 2) an expression for the *effective* number of parameters $p_{eff}(\lambda)$

for nonlinear learning systems (p_{eff} generally differs from the true number of parameters or weights p), 3) an expression for the optimal regularization parameter λ_{opt} for such a system, and 4) the generalized prediction error (GPE) estimate of prediction risk for nonlinear learning systems. These results extend to the general nonlinear case a number of well known results obtained by the statisticians Mallows, Akaike, Barron, and Wahba.

4.2 Architecture Selection and Bond Rating

Closely coupled to this theoretical work is my recent algorithmic work on principled architecture selection methods (Utans and Moody 1991). We proposed a heuristic search procedure for finding a near optimal network architecture within the space of possible architectures for a given problem. The search over architectures determines the number of units in a network, the subset of available variables to use as inputs, and the topology of the network. Estimates of prediction risk are used to choose the optimal architecture within the set of those considered. We have successfully applied this approach to developing a system for predicting corporate bond ratings.